UNIVERSITY OF TORONTO
Faculty of Arts & Science

STA365 W24 Midterm Examination

Wednesday, Feb 14, 2024

Duration: 2 hours.
Aids Allowed: None.
Instructions: Write answers in the space provided in the exam.
Exam Reminders:

- Fill out your name and UTORid on this page and on the scantron answer sheet attached as the last page of the exam.
- Do not begin writing the actual exam until the announcements have ended and the Exam Facilitator has started the exam.
- As a student, you help create a fair and inclusive writing environment. If you possess an unauthorized aid during an exam, you may be charged with an academic offence.
- Turn off and place all cell phones, smart watches, electronic devices, and unauthorized study materials in your bag under your desk. If it is left in your pocket, it may be an academic offence.
- When you are done your exam, raise your hand for someone to come and collect your exam. Do not collect your bag and jacket before your exam is handed in.
- If you are feeling ill and unable to finish your exam, please bring it to the attention of an Exam Facilitator so it can be recorded before leaving the exam hall.
- In the event of a fire alarm, do not check your cell phone when escorted outside.

0. To keep things fair, no questions about the exam will be answered during the duration of the exam. If you think there is a problem with a question, note the question and briefly describe the problem in the space below and your concern will be evaluated during marking.

1. A *Poisson distribution* with $E[X] = \lambda$ models the chance $\Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$ that x events occur in a fixed time period, while as a function of λ this *probability mass function* looks like the *kernel* $\lambda^x e^{-\lambda}$ of a *gamma distribution probability density function* with parameters $\alpha = x + 1$ and $\beta = 1$.

   For a randomized clinical trial setting with equally sized treatment and control groups where the outcomes are occurrences of adverse health events, give posterior distributions for the two groups based on *improper priors* $p(\lambda_k) \propto 1$ which provide a Bayesian analysis framework that could be used to stop the trial at any time that we have evidence leading us to believe that the outcomes in one of the groups have a 95% chance of being better than the other group. *Hint: just give the posteriors.*

   1 point $\lambda_k \sim Gamma(\alpha = x_k + 1, \beta = 1)$ [partial credit may be subjectively awarded]
   [1/2 point for binomial specifications attempting Bayesian bandit-like frameworks]

2. Give pseudocode estimating your belief about the chance that the treatment group has better outcomes than the control group. Clearly specify what the parameters of the posterior distributions in question would be at the time such a determination was made, and what subsequent actions you might take in response to these determinations and why. *Hint: both of the groups are equally sized.*

   1.5 points [partial credit may be subjectively awarded] [1 point for binomial Bayesian bandit versions]

```
from scipy import stats # not required
draws=10000 # any large number is fine
(stats.gamma(alpha=x_contrl+1, beta=1).rvs(draws)>
 stats.gamma(alpha=x_trtmnt+1, beta=1).rvs(draws)).mean() #equivalent "code" is fine
# beta "rate" parameter -- it's not important that the above functions actually work
# Posterior predictive distributions creating Poisson samples are fine/not necessary
```

   0.5 points If greater than 0.95 (95%), stop the trial, we believe the treatment is beneficial and so it should be given to all study participants [partial credit may be subjectively awarded] [not necessary, but it's also the case that we might stop the trial for small probabilities, too...]

3. The "multiple comparisons problem" refers to the fact that any choice to reject a null hypothesis in a hypothesis testing context entails a chance of a Type I error; so, the more tests that are carried out, the greater the chance that there might be a Type I error within the collection of rejected null hypotheses. In a frequentist context, this is often addressed through the use of a "multiple testing correction" such as a *Bonferroni* (or similar) adjustment to control the *family-wise error rate*, or even the conversion of *p-values* to *false discovery rates* (so-called *q-values*) in more extreme circumstances.

   In the adaptive methodology defined in the previous questions, we will repeatedly compare the treatment and control groups. From a frequentist perspective, such repeated comparisons (regardless of their correlation) will necessarily increase the Type I error rate of a final decision to prefer one of the two groups. While this consideration is relevant, this consideration is frequentist rather than Bayesian since Bayesian decisions are not justified in terms of Type I error rate guarantees. What is the Bayesian justification for their decision to prefer the treatment group to the control group?

   1 point Our belief about that chance (the posterior probability) that the treatment intervention is better than the control (on average) [decreasing partial credit for increasingly dissimilar statements]

4. Supposing you have decided to stop the study, provide pseudocode giving an interval estimate conveying the uncertainty in your belief about the chance of avoiding an adverse event over the current duration of the study if you've been given the treatment. State the interpretation of this interval in simple terms. *Hint: assume there are n study participants in the treatment group and consider* $\frac{n-\lambda}{n}$.

2 points [1 point for quantile and 1 point for quantile or percentile function]
[partial credit may be subjectively awarded] [1 point for binomial Bayesian bandit-like versions]

```
import numpy as np; from scipy import stats # not required
np.quantile((n-stats.gamma(alpha=x_trtmnt+1, beta=1).rvs(draws))/n, [0.025, 0.975])
# Any interval is fine, 80%, 90%, etc. as long as it's correctly interpreted below
# and percentile or sorting based code is fine if it's expressing equivalent ideas
# We don't need this/working code but we need to see the right idea clearly expressed
# Posterior predictive distributions creating Poisson samples are fine/not necessary
```

1 point [this needs to be basically correct and matching the probability above for credit]
We believe there's a 95% chance (posterior probability) that the (expected) proportion of adverse events will be between the upper and lower bounds given above.

5. Now provide pseudocode giving interval estimates conveying the uncertainty in your belief about (a) the *relative risk* $\frac{\lambda_C}{\lambda_T}$ and (b) the *absolute risk* $\frac{\lambda_C}{n} - \frac{\lambda_T}{n}$ of the control group compared to the treatment group. *Hint: evaluating* $\lambda_C^{(t)}/\lambda_T^{(t)}$ *is similar to evaluating* $\lambda_T^{(t)} < \lambda_C^{(t)}$ *or just using* $\lambda_T^{(t)}$.

1 point [1/2 point each]
[partial credit may be subjectively awarded] [1/2 point for binomial Bayesian bandit-like versions]

```
import numpy as np; from scipy import stats # not required
np.quantile(stats.gamma(alpha=x_trtmnt+1, beta=1).rvs(draws)/
            stats.gamma(alpha=x_trtmnt+1, beta=1).rvs(draws), [0.025, 0.975])
np.quantile( stats.gamma(alpha=x_trtmnt+1, beta=1).rvs(draws)/n
            -stats.gamma(alpha=x_trtmnt+1, beta=1).rvs(draws)/n, [0.025, 0.975])
# Any interval is fine, 80%, 90%, etc. and percentile or sorting based is fine
# We don't need this/working code but we need to see the right idea clearly expressed
# Posterior predictive distributions creating Poisson samples are fine/not necessary
```

6. Suppose $1 >> \epsilon_{d'} > \epsilon_d > 0$, so ϵ_d and $\epsilon_{d'}$ below are both very small. What is the implication of

$$E_{(\lambda_T,\lambda_C)|\text{data}}[\lambda_C - \lambda_T] \overset{E[\lambda_C/\lambda_T]\approx 1}{=} \epsilon_d \quad \longrightarrow \quad E_{(\lambda'_T,\lambda'_C)|\text{more data}}[\lambda'_C - \lambda'_T] \overset{E[\lambda'_C/\lambda'_T]>2}{=} a$$

$$E_{(\lambda_T,\lambda_C)|\text{data}}[\tfrac{\lambda_C}{n} - \tfrac{\lambda_T}{n}] = \tfrac{\epsilon_d}{n} \quad \longrightarrow \quad E_{(\lambda'_T,\lambda'_C)|\text{more data}}[\tfrac{\lambda'_C}{n} - \tfrac{\lambda'_T}{n}] = \tfrac{a}{n} = \epsilon_{d'}$$

when for $\lambda_C \approx \lambda_T \approx 2$ for some study duration d it's not so improbable to observe $x_T = 6$ and $x_C = 1$ which would lead to the "Type II error" posterior belief $\Pr(\lambda_T > \lambda_C) = \Pr(\frac{\lambda_T}{\lambda_C} > 1) > 0.95$. *Hint: classical* Type II error *is concerned with test* power*, here considered in a low-prevalence context.*

1 point Long running studies are needed in contexts examining low-prevalence events [since in low-prevalence events context we won't expect appropriately actionable observable events unless the study has a sufficiently long duration...]. [decreasing partial credit for increasingly dissimilar statements]

7. Suppose you have the following *normal likelihood* and *gamma prior* for the *precision* parameter ϕ.

$$\left[\prod_{i=1}^{n}\sqrt{\frac{\phi}{2\pi}}e^{-\frac{\phi x_i^2}{2}}\right] \qquad \frac{\beta^\alpha}{\Gamma(\alpha)}\phi^{\alpha-1}e^{-\beta\phi}$$

Show the derivation of the *(hyper)parameters* of the *posterior distribution* of ϕ and the report what type of distribution the *posterior distribution* of ϕ is, justifying in some manner why this must be so.

1 point [for showing work] [partial credit may be subjectively awarded]

$$\left(\frac{\phi}{2\pi}\right)^{\frac{n}{2}} e^{-\frac{\phi\sum_{i=1}^n x_i^2}{2}} \frac{\beta^\alpha}{\Gamma(\alpha)}\phi^{\alpha-1}e^{-\beta\phi} \propto \phi^{\frac{n}{2}} e^{-\frac{\phi\sum_{i=1}^n x_i^2}{2}} \phi^{\alpha-1}e^{-\beta\phi}$$
$$= \phi^{\alpha+\frac{n}{2}-1} e^{-\left(\beta+\frac{\sum_{i=1}^n x_i^2}{2}\right)\phi}$$

0.5 points
so the parameters of the posterior distribution for ϕ are [shape] $\alpha + \frac{n}{2}$ and [rate] $\beta + \frac{\sum_{i=1}^n x_i^2}{2}$

0.5 points
for a *gamma distribution* since this *kernel* could only be that of a *gamma distribution*

8. The *posterior mean* $\frac{(\tau\theta_0+\phi\sum_{i=1}^n x_i)}{(\tau+n\phi)}$ and *posterior precision* $\tau + n\phi$ for the *normal-normal* model

$$\left[\prod_{i=1}^{n}\sqrt{\frac{\phi}{2\pi}}e^{-\frac{\phi(x_i-\theta)^2}{2}}\right]\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau(\theta-\theta_0)^2}{2}}$$

specify how the *prior* and the *likelihood* combine to create the *posterior* (assuming ϕ is known).

- How many "observations" does the *prior* contribute to the *posterior*?
  Hint: the sample contributes n observations. $n_0 = \frac{\tau}{\phi}$ [0.5 points]
- What is the *posterior distribution* analog to the familiar *standard error* formula $\text{Var}[\bar{x}]^{\frac{1}{2}} = \frac{\sigma_x}{\sqrt{n}}$?
  Hint: consider assuming $\tau = n_0\phi$. $\frac{\phi^{-\frac{1}{2}}}{\sqrt{n_0+n}}$ or $\frac{\sigma_x}{\sqrt{n_0+n}}$ [0.5 points]
- In simple language, what is the *posterior mean* in terms of the *prior mean* and *sample mean*?

1 point The weighted average of the prior and sample mean, weighted by their "contributions" n_0 and n, respectively

9. What is the *posterior mean* and *standard deviation* for the previous problem if an *improper prior* $p(\theta) \propto 1$ was used?

   1 point [1/2 point each]

$$\frac{(\cancel{\tau\theta_0} + \cancel{\tau}\phi \sum_{i=1}^n x_i)}{(\cancel{\tau} + n\phi)} = \frac{\cancel{\phi} \sum_{i=1}^n x_i}{n\cancel{\phi}} = \bar{x} \qquad \cancel{\tau} + n\phi = n\phi \longrightarrow \frac{\phi^{-\frac{1}{2}}}{\sqrt{n}} \text{ or } \frac{\sigma_x}{\sqrt{n}}$$

10. What is *autocorrelation* in the MCMC context and how does it affect *effective posterior sample size*?

    2 points [1 point for each sentence] Autocorrelation is the correlation between $\theta^{(t)}$ and $\theta^{(t+k)}$ for some offset k [such as $k = 1$] within a Markov chain caused by the [decaying] sequential dependence of the samples [indexed by t]. The effective samples size is reduced by positive autocorrelations [according to $n_{\text{eff}} = \frac{1}{1+2\sum_{k=1}^{\infty} \rho_k}$]. [decreasing partial credit for increasingly dissimilar statements]

11. Theoretically speaking, a *Markov chain* must converge to a *stationary distribution* before it can be used for MCMC. In simple terms, how can we practically examine empirical concerns about this?

    1 point Multiple Markov chains should be run and potential "lack of agreement" rejected using the split-$\hat{R} < 1.05$ rule of thumb. [-0.5 points for not mentioning split-$\hat{R}$]

12. Does *Hamiltonian Monte Carlo (HMC)* methodology require that we know the *normalizing constant* of a *posterior distribution* in order to use it for MCMC? Justify your answer.

    0.5 points
    No,
    1 point
    HMC is based on Metropolis-Hastings acceptance rates which do not require normalizing constants and students may clarify that $\frac{p(\tilde{\theta}^{(t)}, \tilde{v}^{(t)} + \epsilon | x)}{p(\theta^{(t-1)}, v^{(t-1)} | x)} = \frac{cp(\tilde{\theta}^{(t)}, \tilde{v}^{(t)} + \epsilon | x)}{cp(\theta^{(t-1)}, v^{(t-1)} | x)} = \frac{p(\tilde{\theta}^{(t)}, \tilde{v}^{(t)} + \epsilon, x)}{p(\theta^{(t-1)}, v^{(t-1)}, x)}$ in the MH formula
    0.5 points
    and Hamiltonian dynamics uses log densities so normalizing constants vanish in partial derivatives.

13. Provide a *probabilistic programming* specification to estimate the endpoints of a uniform distribution $U_i \sim U(a, b)$ based on positive "double digit" valued `np.array u` of actualizations of this distribution.

Hint: your code should be "close to working" and only require parameter name related edits to run.

2 points [per comments below]
[Code does not have to be "perfectly working" but should be something like what's given below]
[decreasing partial credit for increasingly dissimilar statements]

```
import pymc as pm
uniform_limits_model = pm.Model()
with uniform_limits_model:
# 1 point for reasonable a and b below
# 1 point for reasonable u below
# -1/2 point if 'x = pm.Dist("x"' structure is not used
    a = pm.Gamma("a", beta=1, alpha=10) # other reasonable options allowable
    b = pm.Gamma("b", beta=1, alpha=50) # other reasonable options allowable
    u = pm.Uniform("u", upper=a, lower=b, obs=u) # Uniform and obs are needed


with uniform_limits_model:
    HMC_idata = pm.sample()
```

14. Explain what *Gibbs sampling* is and describe the *"curse of dimensionality"* that it suffers from.

[partial credit may be subjectively awarded]
0.5 points
Gibbs sampling is something like alternating sampling from $p(\theta|\phi, x)$ and $p(\phi|\theta, x)$, but can extend to more than just two parameters in general by cyclically sampling full conditional distributions.
0.5 points
The "curse of dimensionality" in Gibbs sampling is that it only samples along one axis at a time and cannot change all coordinates simultaneously so it cannot explore a high dimensional space efficiently.

15. Illustrate what a *divergence* is in the *HMC* context and discuss the primary problem these cause and what the drawbacks are of "solutions" which attempt to avoid *divergences* by using "shorter steps".

[partial credit may be subjectively awarded]
0.75 points
[Students should do something like draw a line moving off the contours of a energy function/logpdf]
0.5 points
The main problem is rejected proposals. [Low energy $-\infty$ / low probability 0 proposals (divergences) aren't accepted resulting in a "sticky chain" or requiring shorter time steps, resulting in either increased autocorrelation in Markov chains or increased computational costs...]
0.75 points
Shorter steps means more steps (and hence more computations) are needed for equivalent movement "away" from current values; thus, making it more computationally expensive to produce reduced autocorrelation Markov chains while still not too frequently encountering divergences.
[decreasing partial credit for increasingly dissimilar statements]

16. Show that as a function of β the *multiple linear regression model* specification

$$(2\pi)^{-n/2}\det(\mathbf{\Sigma})^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^\top\mathbf{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\beta)\right)$$

is proportional to the *multivariate normal distribution*

$$\mathcal{MVN}(E[\beta|\mathbf{X},\mathbf{y}.\mathbf{\Sigma}]=\text{Var}[\beta|\mathbf{X},\mathbf{y}.\mathbf{\Sigma}]^{-1}\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{y},\text{Var}[\beta|\mathbf{X},\mathbf{y}.\mathbf{\Sigma}]=\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{X})$$

and write down the density of the *LKJ prior* for *correlation matrix* $\mathbf{R}$. 0.5 points $p(\mathbf{R})\propto\det(\mathbf{R})^{\eta-1}$

1 point (for sufficiently showing the core relationship)

$$\begin{aligned}
&\exp\left((\mathbf{y}-\mathbf{X}\beta)^\top\mathbf{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\beta)\right)\\
\propto\ &\exp\left((\mathbf{X}\beta)^\top\mathbf{\Sigma}^{-1}(\mathbf{X}\beta)-2(\mathbf{X}\beta)^\top\Sigma^{-1}\mathbf{y}\right)\\
=\ &\exp\left(\beta^\top(\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{X})\beta-2(\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{X})(\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}(\beta^\top\mathbf{X}^\top\Sigma^{-1}\mathbf{y}\right)\\
\propto\ &\exp\left((\beta-(\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathbf{y})^\top\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{X}(\beta-(\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathbf{y})\right)
\end{aligned}$$

0.5 points (for making the correct conlusion)
Multivariate normal distribution kernel above shows distributional parameters:

$$\exp\left((\beta-\underbrace{(\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathbf{y}}_{E[\beta|\mathbf{X},\mathbf{y}.\mathbf{\Sigma}]})^\top\underbrace{\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{X}}_{\text{Var}[\beta|\mathbf{X},\mathbf{y}.\mathbf{\Sigma}]}(\beta-(\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Sigma^{-1}\mathbf{y})\right)$$

17. General *determinant* and *inversion* computations are $O(p^3)$ which means that their computation for $p\times p$ matrices require a number of operations that is proportional to p^3. Using *Cholesky* factorizations $\det(\mathbf{\Sigma})=\det(\mathbf{L}\mathbf{L}^\mathbf{T})=\prod_{k=1}^p\mathbf{L}_{kk}^2$ and $(\mathbf{y}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})=\boldsymbol{\epsilon}^\top\mathbf{L}^{-\top}\mathbf{L}^{-1}\boldsymbol{\epsilon}\underbrace{=(\mathbf{L}^{-1}\boldsymbol{\epsilon})^\top(\mathbf{L}^{-1}\boldsymbol{\epsilon})=\mathbf{x}^\top\mathbf{x}}_{\text{solving for }\mathbf{x}\text{ in }\mathbf{L}\mathbf{x}=\boldsymbol{\epsilon}\text{ is }O(n^2)}$.
How could this benefit *HMC* even if the *posterior distribution* of interest is not *normally distributed*?

1 point (1/2 for likelihood or v and +1/4 for both and 1/4 for higher dimension relevance)
Data models might include multivariate normal contributions to the likelihood [which would benefit from both computational shortcuts above] even if the posterior is not multivariate normal; and, the auxiliary parameters v in HMC could be specified as an multivariate normal distribution [independent of θ] so the *kinetic energy* $K(v)$ contribution could also benefit from the ["solve for $\mathbf{x}$"] computational shortcut. The benefit would be increasingly noticeable in higher dimensions, where HMC is most useful...

18. When would *HMC* have either a "sticky" chain or a "poorly mixing" chain?

    1 point (1/2 point for each)
    Sticky: steps too long causing divergences / very low acceptance rates
    Poorly mixing: steps too short, causing high autocorrelation in chains
    [so these are the extreme behaviours of too long or too short step sizes]

19. For *Bernoulli* $\tilde{p}(\tilde{X} = \tilde{x}) = \frac{1}{2}^{\tilde{x}} \frac{1}{2}^{1-\tilde{x}}$ and $p(X = x) = \frac{1}{3}^{x} \frac{2}{3}^{1-x}$ and $\tilde{p}(\tilde{X} = \tilde{x})$ the proposal distribution for a *Metropolis-Hastings* algorithm sampling from the stationary distribution $p(X = x)$, what are the transition probabilities

    - $Pr(X^{(t)} = 1 | x^{(t-1)} = 1) = 1 \times \frac{1}{2} = \frac{1}{2}$
    - $Pr(X^{(t)} = 0 | x^{(t-1)} = 1) = 1 \times \frac{1}{2} = \frac{1}{2}$
    - $Pr(X^{(t)} = 1 | x^{(t-1)} = 0) = \frac{1/3}{2/3} \times \frac{1}{2} = \frac{1}{4}$
    - $Pr(X^{(t)} = 0 | x^{(t-1)} = 0) = \frac{1}{4} + 1 \times \frac{1}{2} = \frac{3}{4}$

    1 point (1/4 point for each above)
    Chance of proposing 0 or 1 is 50/50 needs to be included.
    Higher or equal density is always accepted in the MH proposal acceptance step.
    Proposal is symmetric so not needed in acceptance step probability calculations.
    If a proposal is rejected with some probability, that probability goes to the option keeping the previous value, which is what's happening in the final calculation.

20. The sequence $1, 0, 0, 0, 0, 1$ satisfies the expected behaviour of $p(X = x)$ from the previous problem. Using the transition probabilities from the previous problem, (a) describe how this sequence plausibly came about as a *Markov chain* from the specified *Metropolis-Hastings* algorithm, and (b) demonstrate that, if the first two numbers proportionally represent sampling that occurs during the $x^{(t-1)} = 1$ state of the algorithm which happens one-third of the time, and the last four numbers proportionally represent sampling that occurs during the $x^{(t-1)} = 0$ state of the algorithm that happens two-thirds of the time, then the "long term proportionality" behaviour of the *Markov chain* will indeed proportionally coincides with sampling from $p(X = x)$.

    1 point (1/4 point for each part below)
    (a) An initial state of 1 has a 50/50 shot at becoming a 0 or a 1; so, it might become a 0. [(b) The first two values 0 and 1 then reflect the long term 50/50 behaviour contributed by the $x^{(t-1)} = 0$ state of the algorithm which occurs one-third of the time.] (a) The subsequent state of 0 has a 1/4 chance of drawing 1; so, it might do so on it's fourth attempt. [(b) The last three 0's and 1 then reflect the long term 75%/25% behaviour contributed by the $x^{(t-1)} = 1$ state of the algorithm which occurs two-thirds of the time.] (b) So the long-term chance of 1 is $\frac{1}{3} \times \frac{1}{2} + \frac{2}{3} \times \frac{1}{4} = \frac{1}{6} + \frac{2}{12} = \frac{1}{3}$ while the long-term chance of 0 is $\frac{1}{3} \times \frac{1}{2} + \frac{2}{3} \times \frac{3}{4} = \frac{1}{6} + \frac{1}{2} = \frac{2}{3}$ which reflects $p(X = x)$.